Asymmetric Loss For Multi-Label Classification

Tal Ridnik* Emanuel Ben-Baruch*
Nadav Zamir Asaf Noy Itamar Friedman Matan Protter Lihi Zelnik-Manor

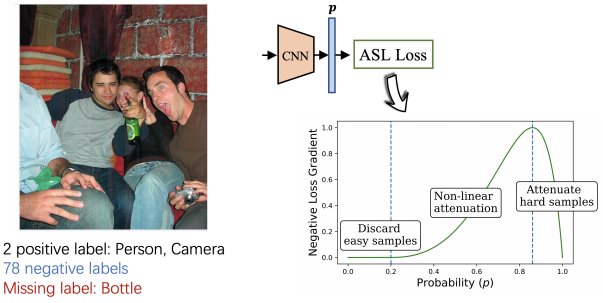
DAMO Academy, Alibaba Group
{emanuel.benbaruch, tal.ridnik, nadav.zamir, asaf.noy, itamar.friedman, matan.protter, lihi.zelnik}
@alibaba-inc.com

Abstract

In a typical multi-label setting, a picture contains on average few positive labels, and many negative ones. This positive-negative imbalance dominates the optimization process, and can lead to under-emphasizing gradients from positive labels during training, resulting in poor accuracy. In this paper, we introduce a novel asymmetric loss ("ASL"), which operates differently on positive and negative samples. The loss enables to dynamically down-weights and hard-thresholds easy negative samples, while also discarding possibly mislabeled samples. We demonstrate how ASL can balance the probabilities of different samples, and how this balancing is translated to better mAP scores. With ASL, we reach state-of-the-art results on multiple popular multi-label datasets: MS-COCO, Pascal-VOC, NUS-WIDE and Open Images. We also demonstrate ASL applicability for other tasks, such as single-label classification and object detection. ASL is effective, easy to implement, and does not increase the training time or complexity. Implementation is available at: https://github.com/Alibaba-MIIL/ASL.

1. Introduction

Typical natural images contain multiple objects and concepts [34, 37], highlighting the importance of multi-label classification for real-world tasks. Recently, remarkable advances have been made in multi-label benchmarks such as MS-COCO [20], NUS-WIDE [7], Pascal-VOC [11] and Open Images [17]. Notable success was reported by exploiting label correlation via graph neural networks which represent the label relationships [6, 5, 10] or word embeddings based on knowledge priors [6, 31]. Other approaches are based on modeling image parts and attentional regions [36, 12, 32, 35], as well as using recurrent neural networks [22, 29].



(a) (b)

Figure 1: (a) **Real world challenges in multi-label classification**. A typical image contains few positive samples, and many negative ones, leading to high negative-positive imbalance. Also, missing labels in ground-truth are common in multi-label datasets. (b) **Proposed solution with ASL**. The loss properties will be detailed in Section 2.5

Despite their effectiveness, recent approaches are characterized by extensive architecture modifications and relying on additional external information, such as word embeddings and NLP models. In this work, we question whether such intricate solutions are truly necessary for achieving high performance in multi-label classification tasks. In particular, we demonstrate that a careful design of the loss function can greatly benefit classification accuracy, while still maintaining a simple and efficient solution, based on standard architectures and training schemes.

A key characteristic of multi-label classification is the inherent positive-negative imbalance created when the overall number of labels is large. Most images contain only a small fraction of the possible labels, implying that the number of positive samples per category will be, on average, much lower than the number of negative samples. To address this, [33] suggested a loss function for statically handling the imbalance in multi-label problems. However, it was aimed

*Equal contribution

specifically at long-tail distribution scenarios. High imbalance is also encountered in dense object detection, where it stems from the ratio of foreground vs. background regions. Some solutions based on resampling methods were proposed, by selecting only a subset of the possible background examples [23]. However, resampling methods are not suitable for handling multi-label classification imbalancing, since each image contains many labels, and resampling cannot change the distribution of only a specific label.

Another common solution in object detection is to adopt the focal loss [19], which decays the loss as the label's confidence increases. This puts focus on hard samples, while down-weighting easy samples, which are mostly related to easy background locations. Surprisingly, focal loss is seldom used for multi-label classification, and cross-entropy is often the default choice (see [6, 1, 4, 21, 12], for example). Since high negative-positive imbalance is also encountered in multi-label classification, focal loss might provide better results, as it encourages focusing on relevant hard-negative samples, which are mostly related to images that do not contain the positive class, but do contain some other confusing categories. Nevertheless, for the case of multi-label classification, treating the positive and negative samples equally, as proposed by focal loss, is sub-optimal, as it results in the accumulation of more loss gradients from negative samples, and down-weighting of important contributions from the rare positive samples. In other words, the network might focus on learning features from negative samples while underemphasizing learning features from positive samples.

In this paper, we introduce an asymmetric loss (ASL) for multi-label classification, which explicitly addresses the negative-positive imbalance. ASL is based on two key properties: first, to focus on hard negatives while maintaining the contribution of positive samples, we decouple the modulations of the positive and negative samples and assign them different exponential decay factors. Second, we propose to shift the probabilities of negative samples to completely discard very easy negatives (hard thresholding). By formulating the loss derivatives, we demonstrate that probability shifting also enables to discard very hard negative samples, suspected as mislabeled, which are common in multi-label problems [10].

We compare ASL to the common symmetrical loss functions, cross-entropy and focal loss, and show significant mAP improvement using our asymmetrical formulation. By analyzing the model's probabilities, we demonstrate the effectiveness of ASL in balancing between negative and positive samples. We also introduce a method that dynamically adjusts the asymmetry level throughout the training process, by demanding a fixed gap between positive and negative average probabilities, allowing simplification of the hyper-parameter selection process.

The paper's contributions can be summarized as follow:

- We design a novel loss function, ASL, which explicitly copes with two main challenges in multi-label classification: high negative-positive imbalance, and ground-truth mislabeling.
- We thoroughly study the loss properties via detailed gradient analysis. An adaptive procedure for controlling the asymmetry level of the loss is introduced, to simplify the process of hyper-parameter selection.
- Using ASL, we obtain state-of-the-art results on four popular multi-label benchmarks. For example, we reach 86.6% mAP on MS-COCO dataset, surpassing the previous top result by 2.8%.
- Our solution is effective and easy to use. It is based on standard architectures, does not increase training and inference time, and does not need any external information, in contrast to recent approaches. To make ASL accessible, we will share our trained models and a fully reproducible training code.

2. Asymmetric Loss

In this section, we will first review cross-entropy and focal loss. Then we will introduce the components of the proposed asymmetric loss (ASL), designed to address the inherent imbalance nature of multi-label datasets. We will also analyze ASL gradients, provide probability analysis, and present a method to set the loss' asymmetry levels during training dynamically.

2.1. Binary Cross-Entropy and Focal Loss

As commonly done in multi-label classification, we reduce the problem to a series of binary classification tasks. Given K labels, the base network outputs one logit per label, z_k. Each logit is independently activated by a sigmoid function $\sigma(z_k)$. Let's denote y_k as the ground-truth for class k. The total classification loss, L_{tot}, is obtained by aggregating a binary loss from K labels:

$$L_{\text{tot}} = \sum_{k=1}^{K} L\left(\sigma(z_k), y_k\right). \tag{1}$$

A general form of a binary loss per label, L, is given by:

$$L = -yL_{+} - (1-y)L_{-} \tag{2}$$

Where y is the ground-truth label (for brevity we omitted the class index k), and L_{+} and L_{-} are the positive and negative loss parts, respectively. Following [19], focal loss is obtained by setting L_{+} and L_{-} as:

$$\begin{cases} L_{+} = (1-p)^{\gamma}\log(p) \\ \\ L_{-} = p^{\gamma}\log(1-p) \end{cases} \tag{3}$$

where $p = \sigma(z)$ is the network's output probability and γ is the *focusing parameter*. $\gamma = 0$ yields binary cross-entropy.

By setting $\gamma > 0$ in Eq. 3, the contribution of easy negatives (having low probability, $p \ll 0.5$) can be down-weighted in the loss function, enabling to focus more on harder samples during training.

2.2. Asymmetric Focusing

When using focal loss for multi-label training, there is an inner trade-off: setting high γ, to sufficiently down-weight the contribution from easy negatives, may eliminate the gradients from the rare positive samples. We propose to decouple the focusing levels of the positive and negative samples. Let γ_+ and γ_- be the positive and negative focusing parameters, respectively. We obtain asymmetric focusing by re-defining the loss:

$$\begin{cases} L_+ = (1-p)^{\gamma_+}\log(p) \\ \\ L_- = p^{\gamma_-}\log(1-p) \end{cases} \tag{4}$$

Since we are interested in emphasizing the contribution of positive samples, we usually set $\gamma_- > \gamma_+$. Asymmetric focusing decouples the decay rates of positive and negative samples. Through this, we achieve better control over the contribution of positive and negative samples to the loss function, and help the network learn meaningful features from positive samples, despite their rarity.

It should be noted that methods which address class imbalance via static weighting factors were proposed in previous works [15, 9]. However, [19] found that those weighting factors interact with the focusing parameter, making it necessary to select the two together. In practice, [19] even suggested a weighting factor which favors background samples ($\alpha = 0.25$). In section 3 we will show that simple linear weighting is insufficient to tackle the negative-positive imbalance issue in multi-label classification properly. For those reasons, we chose to avoid adding static weighting factors to our focusing formulation.

2.3. Asymmetric Probability Shifting

Asymmetric focusing reduces the contribution of negative samples to the loss when their probability is low (soft thresholding). Since the level of imbalancing in multi-label classification can be very high, this attenuation is not always sufficient. Hence, we propose an additional asymmetric mechanism, probability shifting, that performs hard thresholding of very easy negative samples, i.e., it fully discards negative samples when their probability is very low. Let's define the *shifted probability*, p_m, as:

$$p_m = \max(p - m, 0) \tag{5}$$

where the *probability margin* $m \geq 0$ is a tunable hyper-parameter. Integrating p_m into L_- of Eq.(3), we get an asymmetric probability-shifted focal loss:

$$L_- = (p_m)^{\gamma}\log(1-p_m) \tag{6}$$

In Figure 2 we draw the probability-shifted focal loss, for negative samples, and compare it to regular focal loss and cross-entropy. From a geometrical point-of-view, we can

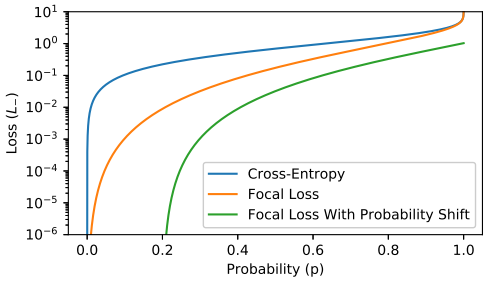


Figure 2: **Loss Comparisons.** Comparing probability-shifted focal loss to regular focal loss and cross-entropy, for negative samples. We used $\gamma_- = 2$ and $m = 0.2$.

see that probability shifting is equivalent to moving the loss function to the right, by a factor m, thus getting $L_- = 0$ when $p < m$. We will later show, via gradient analysis, another important property of the probability shifting mechanism - it can also reject mislabeled negative samples.

Notice that the concept of probability shifting is not limited to cross-entropy or focal loss, and can be used on many loss functions. Linear hinge loss [1], for example, can also be seen as (symmetric) probability shifting of linear loss. Also notice that logits shifting, as suggested in [19] and [33], is different from probability shifting due to the non-linear sigmoid operation.

2.4. ASL Definition

To define the Asymmetric Loss (ASL), we integrate the two mechanisms of asymmetric focusing and probability shifting into a unified formula:

$$ASL = \begin{cases} L_+ = & (1-p)^{\gamma_+}\log(p) \\ \\ L_- = & (p_m)^{\gamma_-}\log(1-p_m) \end{cases} \tag{7}$$

Where p_m is defined in Eq.(5). ASL allows us to apply two types of asymmetry for reducing the contribution of easy negative samples to the loss function - soft thresholding via the focusing parameters $\gamma_- > \gamma_+$, and hard thresholding via the probability margin m.

It can be convenient to set $\gamma_+ = 0$, so that positive samples will incur simple cross-entropy loss, and control the level of asymmetric focusing via a single hyper-parameter, γ_-. For experimentation and generalizability, we still keep the γ_+ degree of freedom.

2.5. Gradient Analysis

To better understand the properties and behavior of ASL, we next provide an analysis of the loss gradients, in comparison to the gradients of cross entropy and focal loss. Looking at the gradients is useful since, in practice, the network weights are updated according to the gradient of the loss, with respect to the input logit z. The loss gradients for negative samples in ASL are:

$$\begin{aligned}\frac{dL_-}{dz} &= \frac{\partial L_-}{\partial p}\frac{\partial p}{\partial z} \\ &= (p_m)^{\gamma_-}\left[\frac{1}{1-p_m} - \frac{\gamma_- \log\left(1-p_m\right)}{p_m}\right]p(1-p)\end{aligned} \tag{8}$$

Where $p = \frac{1}{1+e^{-z}}$, and p_m is defined in Eq.(5). In Figure 3 we present the normalized gradients of ASL, and compare it to other losses. Following Figure 3, we can roughly split the negative samples in ASL into three loss-regimes:

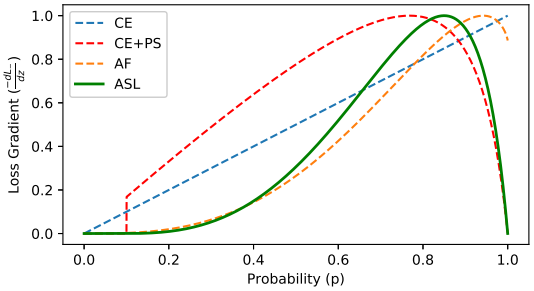


Figure 3: **Gradient Analysis.** Comparing the loss gradients vs. probability for different loss regimes. CE = Cross-Entropy ($m = \gamma_- = 0$), CE+PS = Cross-Entropy with Probability Shifting ($m > 0, \gamma_- = 0$), AF = Asymmetric Focusing ($m = 0, \gamma_- > 0$), ASL ($m > 0, \gamma_- > 0$).

1. Hard-threshold - very easy negatives, with $p < m$, that should be ignored, in order to focus on harder samples.
2. Soft-threshold - negative samples, with $p > m$, that should be attenuated when their probability is low.
3. Mislabeled - very hard negative samples, with $p > p*$, where $p*$ is defined as the point where $\frac{\mathrm{d}}{\mathrm{d}p}\left(\frac{\mathrm{d}L}{\mathrm{d}z}\right) = 0$, which are suspected as mislabeled - when the network computes a very large probability for a negative sample, it is possible that the sample was mislabeled, and its correct label should be positive. It has been shown by [10] that multi-label datasets are prone to mislabeling of negative samples, probably because the manual labeling task is difficult. When dealing with highly imbalanced datasets, even a small mislabeling rate of negative samples largely impact the training. Hence, rejection of mislabelled samples can be beneficial. The rejection must be done carefully, to allow the network to propagate gradients from actual misclassified negative examples.

In Table 1 we compare the properties and abilities of ASL to other losses, according to the gradient analysis. We can

	Hard Threshold	Soft Threshold	Discard Mislabeled	Continuous Gradients
CE	-	-	-	+
AF	-	+	-	+
CE+PS	+	-	+	-
ASL (AF+PS)	+	+	+	+

Table 1: **Properties of different loss** - CE (Cross-Entropy), AF (Asymmetric Focusing), PS (Probability Shifting).

see that only when we combine the two asymmetry mechanisms, focusing and probability margin, we enjoy all the abilities and advantages which are beneficial for imbalanced datasets: hard thresholding of very easy samples, non-linear attenuation of easy samples, rejection of mislabeled samples and continuous loss gradients.

2.6. Probability Analysis

In this section, we wish to provide further support to our claim that in multi-label datasets, using a symmetric loss such as cross entropy or focal loss is sub-optimal for learning positive samples' features. We do that by monitoring the average probabilities outputted by the network during the training. This allows us to evaluate the network's level of confidence for positive and negative samples. Low confidence suggests that features were not learned properly. We begin by defining p_t as:

$$p_t = \begin{cases} \bar{p} & \text{if } y = 1 \\ 1-\bar{p} & \text{otherwise} \end{cases} \tag{9}$$

where $\bar{p}$ denotes the average probability of the samples in a batch at each iteration. Denote by p_t^+ and p_t^- the average probabilities of the positive and negative samples, respectively, and by Δp the probability gap:

$$\Delta p = p_t^+ - p_t^-. \tag{10}$$

A balanced training should demonstrate similar level of mean confidence for positive and negative samples, i.e., Δp should be small throughout and at the end of the training.

In Figure 4 we present the average probabilities p_t^+ and p_t^- along the training, for three different loss functions: cross-entropy, focal loss and ASL. Figure 4 demonstrates the limitation of using symmetric losses for imbalanced datasets. When training with either cross-entropy loss or focal loss, we observe that $p_t^- \gg p_t^+$ (at the end of the training, $\Delta p = -0.23$ and $\Delta p = -0.1$, respectively). This implies that the optimization process gave too much weight to negative samples. Conversely, when training with ASL we can eliminate the gap, implying that the network has the ability to properly emphasize positive samples.

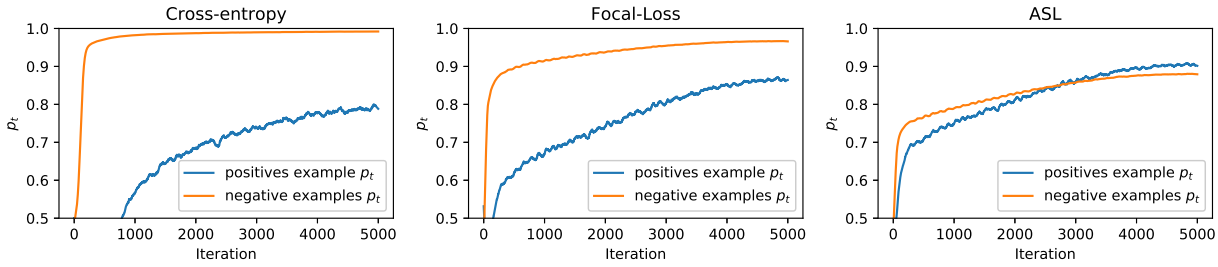


Figure 4: **Probability analysis.** The mean probability of positive and negative samples along the training with cross-entropy, focal loss and ASL, on MS-COCO. For focal loss we used $\gamma = 2$. For ASL we used $\gamma_+ = 0$, $\gamma_- = 2$, $m = 0.2$.

Notice that by lowering the decision threshold p_{th} at inference time (a sample will be declared as positive if $p > p_{th}$), we can control the precision vs. recall trade-off, and favor high true-positive rate over low false-negative rate. However, a large negative probability gap, as obtained by the symmetric losses, suggests that the network under-emphasized gradients from positive samples and converged to a local minima, with sub-optimal performances. We will validate this claim in Section 3.

2.7. Adaptive Asymmetry

Hyper-parameters of a loss function are usually adjusted via a manual tuning process. This process is often cumbersome and requires a level of human expertise. Based on our probability analysis, we wish to offer a simple intuitive way of dynamically adjusting ASL's asymmetry levels, with a single interpretable control parameter.

In the last section we demonstrated that ASL enables to balance a network, and prevent a situation where negative samples have significantly larger p_t than positive samples ($\Delta p < 0$). We now wish to go the other way around, and adjust γ_- dynamically throughout the training, to match a desired probability gap, denoted by Δp_{target}. We can achieve this by a simple adaptation of γ_- after each batch, as described in Eq. 11.

$$\gamma_- \leftarrow \gamma_- + \lambda(\Delta p - \Delta p_{\text{target}}) \tag{11}$$

where λ is a dedicated step size. As we increase Δp_{target}, Eq. 11 enables us to dynamically increase the asymmetry level throughout the training, forcing the optimization process to focus more on the positive samples' gradients. Notice that using similar logic to Eq. 11, we can also dynamically adjust the probability margin, or simultaneously adjust both asymmetry mechanisms. For simplicity, we chose to explore the case of adjusting only γ_- throughout the training, with $\gamma_+ = 0$ and a small fixed probability margin.

Figure 9 in appendix A presents the values of γ_- and Δp throughout the training, for $\Delta p_{\text{target}} = 0.1$. After 10% of the training, the network converges successfully to the target probability gap, and to a stable value of γ_-. In the next section we will analyze the mAP score and possible use-cases for this dynamic scheme.

3. Experimental Study

In this section, we will provide thorough experimentations to better understand the different losses, and demonstrate the improvement we gain from ASL, compared to other losses. We will also test our adaptive asymmetry mechanism, and compare it to a fixed scheme. For testing, we will use the well-known MS-COCO [20] dataset (see Section 4.1.1 for full dataset and training details).

Focal Loss Vs Cross-Entropy: In Figure 5 we present the mAP scores obtained for different values of focal loss γ ($\gamma = 0$ is cross-entropy). We can see from Figure 5 that

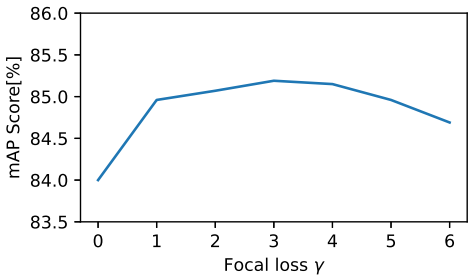


Figure 5: **mAP Vs. Focal Loss γ.** Comparing MS-COCO mAP score for different values of focal loss γ.

with cross-entropy loss, the mAP score is lower than the one obtained with focal loss (84.0% vs 85.1%). Top scores with focal loss are obtained for $2 \leq \gamma \leq 4$. With γ below that range, the loss does not provide enough down-weighting for easy negative samples. With γ above that range, there is too much down-weighting of the rare positive samples.

Asymmetric Focusing: In Figure 6 we test the asymmetric focusing mechanism: for two fixed values of γ_-, 2 and 4, we present the mAP score along the γ_+ axis. Figure 6 demonstrates the effectiveness of asymmetrical focusing - as we decrease γ_+ (hence increasing the level of asymmetry), the mAP score significantly improves.

Interestingly, simply setting $\gamma_+ = 0$ leads to the best re-

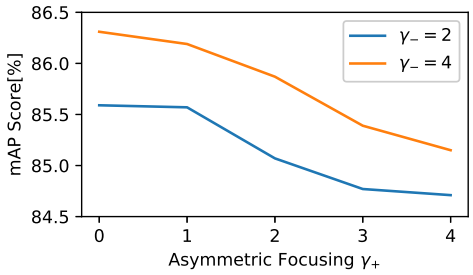


Figure 6: **mAP Vs. Asymmetric Focusing γ_+.** Comparing MS-COCO mAP score for different value of asymmetric focusing γ_+, for $\gamma_- = 2$ and $\gamma_- = 4$.

sults in our experiments. That may further support the importance of keeping the gradient magnitudes high for positive samples. Indeed, allowing $\gamma_+ > 0$ may be useful for cases where there is also an abundance of easy positive samples. Note that we also tried training with $\gamma_+ < 0$, to extend the asymmetry further. However, these trials did not converge, therefore they are not presented in Figure 6.

Asymmetric Probability Margin: In Figure 7 we apply our second asymmetry mechanism, asymmetric probability margin, on top of cross-entropy loss ($\gamma = 0$) and two levels of (symmetric) focal loss, $\gamma = 2$ and $\gamma = 4$.

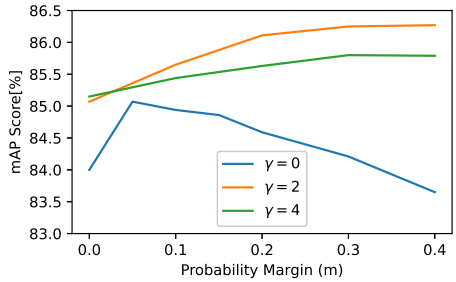


Figure 7: **mAP Vs. Asymmetric Probability Margin.** Comparing MS-COCO mAP score for different values of asymmetric probability margin, on top of a symmetric focal loss, with $\gamma = 0, 2, 4$

We can see from Figure 7 that both for cross-entropy and focal loss, introducing asymmetric probability margin improves the mAP score. For cross-entropy, the optimal probability margin is low, $m = 0.05$, in agreement with our gradient analysis - cross-entropy with probability margin produces a non-smooth loss gradient, with less attenuation of easy samples. Hence, a small probability margin, which still enables hard threshold for very easy samples and rejection of mislabeled samples, is better. For focal loss, the optimal probability margin is significantly higher, $0.3 \leq m \leq 0.4$. This again can be explained by analyzing the loss gradients: since focal loss already has non-linear attenuation of easy samples, we need a larger probability margin to introduce meaningful asymmetry. We can also see that when introducing asymmetric probability margin, better scores are obtained for $\gamma = 2$ compared to $\gamma = 4$, meaning that asymmetric probability margin works better on top of a modest amount of focal loss.

Comparing Different Asymmetries: Until now we tested each ASL asymmetry separately. In Table 2 we present the mAP scores achieved when combining the asymmetries, and compare them to the scores obtained when applying each asymmetry alone. Also, we compare ASL results to another asymmetric mechanism - focal loss combined with linear weighting, as proposed in [23], that statically favors positive samples. The optimal static weight was searched over a range of values between 0.5 to 0.95, with skips of 0.05.

Method	mAP [%]
FL	85.1
FL + linear weighting	85.3
Focusing (ASL)	86.3
Probability margin (ASL)	86.3
Focusing + Probability margin (ASL)	**86.6**

Table 2: **MS-COCO mAP scores for different asymmetric methods.** Focusing mAP obtained for $\gamma_+ = 0, \gamma_- = 3$. Margin mAP obtained for $\gamma = 2, m = 0.3$. Combined mAP obtained for $\gamma_+ = 0, \gamma_- = 4, m = 0.05$.

We can see from Table 2 that the best results are obtained when combining the two components of asymmetry. This correlates with our analysis of the loss gradients in Figure 3, where we demonstrate how combining the two asymmetries enables discarding of very easy samples, nonlinear attenuation of easy samples and rejection of possibly mislabeled very hard negative samples, a result which is not possible when applying only one type of asymmetry. Table 2 also shows that using static weighing is insufficient to properly handle the high negative-positive imbalance in multi-label classification, and ASL, which operates dynamically on easy and hard samples, performs better.

Adaptive Asymmetry: We now examine the effectiveness of adjusting the ASL asymmetry levels dynamically, via the procedure proposed in Eq.11. In Table 3 we present the mAP score, and the final value of γ_-, obtained for various values of Δp_{target}.

We can see from Table 3 that even without any tuning, demanding the unbiased case $\Delta p_{\text{target}} = 0$, a significant improvement is achieved compared to focal loss (85.8% vs. 85.1%). Even better scores are obtained when using a higher probability gap, $\Delta p_{\text{target}} = 0.2$. Interestingly, extra focus on the rare positive samples ($\Delta p_{\text{target}} > 0$) is better than just demanding the unbiased case.

Notice that the top mAP scores obtained from the dy-

Δp_{target}	γ_- Final	mAP Score [%]
0	1.2	85.8
0.1	3.3	86.1
0.2	5.2	**86.4**
0.3	6.2	86.3

Table 3: **Adaptive Asymmetry**. mAP scores and γ_- obtained from adaptive asymmetry runs, for different Δp_{target}.

namic scheme are still lower by 0.2% compared to the best ASL score with a fixed γ_-. One possible reason for this (small) degradation is that the training process is highly impacted by the first epochs [13]. Tuning hyper-parameter dynamically may be sub-optimal at the beginning of the training, which decreases the overall performance. To compensate for the initial recovery iterations, dynamically-tuned γ_- tends to converge to higher values, but the overall score is still somewhat hindered. Due to this decline, we chose to use a fixed asymmetry scheme in section 4.

Still, the dynamic scheme can be appealing to a non-expert user, as it allows control of the asymmetry level via one simple interpretable hyper-parameter. In addition, we will explore in the future ways to expand this scheme for other applications, such as tuning γ_- adaptively per class, which can be impractical with a regular exhaustive search.

4. Dataset Results

In this section, we will evaluate ASL on four popular multi-label classification datasets, and compare its results to known state-of-the-art techniques, and to other commonly used loss functions. We will also test ASL's applicability to other computer vision tasks, such as single-label classification and object detection.

4.1. Multi-Label Datasets

4.1.1 MS-COCO

MS-COCO [20] is a widely used dataset to evaluate computer vision tasks such as object detection, semantic segmentation and image captioning, and has been adopted recently to evaluate multi-label image classification. For multi-label classification, it contains $122,218$ images with 80 different categories, where every image contains on average 2.9 labels, thus giving an average positive-negative ratio of: $\frac{2.9}{80-2.9} = 0.0376$. The dataset is divided to a training set of $82,081$ images and a validation set of $40,137$ images. Following conventional settings for MS-COCO [31, 21], we report the following statistics: mean average precision (mAP), average per-class precision (CP), recall (CR), F1 (CF1) and the average overall precision (OP), recall (OR) and F1 (OF1), for the overall statistics and top-3 highest scores. Among these metrics, mAP, OF1, and CF1 are the main metrics, since they take into account both false-negative and false-positive rates.

In Table 4 we compare ASL results to known state-of-the-art methods from the literature, for the main metrics (Full training details and loss hyper-parameters are provided in appendix B). In Table 7 in appendix C we bring results for all the metrices. We can see from Table 4 that using ASL we significantly outperform previous state-of-the-art methods on ResNet101, the commonly used architecture in multi-label classification, and improve the top mAP score by more than 1%. Other metrics also show improvement.

Method	mAP	CF1	OF1
CADM [5]	82.3	77.0	79.6
ML-GCN [6]	83.0	78.0	80.3
KSSNet [21]	83.7	77.2	81.5
MS-CMA [36]	83.8	78.4	81.0
MCAR [12]	83.8	78.0	80.3
ASL (ResNet101)	**85.0**	**80.3**	**82.3**
ASL (TResNet-L)	**86.6**	**81.4**	**81.8**

Table 4: **Comparison of ASL to state-of-the-art methods on MS-COCO.** All metrics are in %. Results are reported for input resolution 448.

Notice that our ASL-based solution does not require architecture modifications, and does not increase inference and training times. This is in contrast to previous top solutions, which include intricate architecture modifications (attentional regions [12], GCNs [38]), injecting external data like label embeddings [36, 5], and using teacher models [21]. However, ASL is fully complementary to those methods, and employing them as well could lead to further score improvement, at the cost of increasing training complexity and reducing throughput. In addition, we see from Table 4 that using a newer architecture like TResNet-L, that was designed to match the GPU throughput of ResNet101 [25], we can further improve the mAP score, while still keeping the same training and inference time. This is another contribution of our proposed solution - identifying that modern fast architectures can give a big boost to multi-label classification, and the common usage of ResNet101 can be sub-optimal.

In Figure 8 we test the applicability of ASL for different backbones, by comparing the different loss functions on three commonly used architectures: OFA-595 [2], ResNet101 and TResNet-L. We can see from Figure 8 that on all backbones, ASL outperforms focal loss and cross-entropy, demonstrating its robustness to backbone selection, and its superiority over previous loss functions.

Impact of pretraining and input resolutions: In Table 5 we compare the mAP results obtained with a standard ImageNet-1K pretraining, and the newer ImageNet-21K

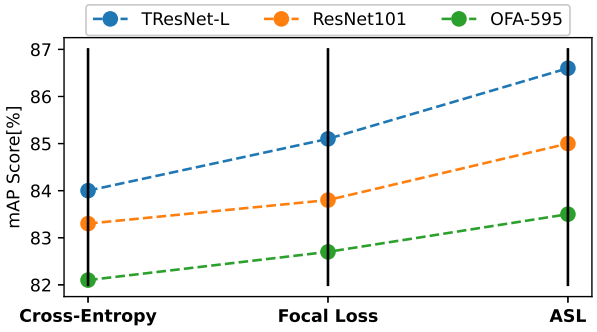


Figure 8: **Testing different losses on various backbones.**

pretraining [24]. We can see that using better pretraining has dramatic impact on the results, increasing the mAP score by almost 2%. We also show in Table 5 that increasing input resolution from 448 to 640 can further improve results.

Method	Architecture	Pretrain Type	Input Resolution	mAP
ASL	TResNet-L	1K	448	86.6
ASL	TResNet-L	21K [24]	448	88.4
ASL	TResNet-L	21K [24]	640	**89.8**

Table 5: **Comparison of MS-COCO mAP scores for different ImageNet pretraining schemes, and input resolutions.** All metrics are in %.

4.1.2 Pascal-VOC

Pascal Visual Object Classes Challenge (VOC 2007) [11] is another popular dataset for multi-label recognition. It contains images from 20 object categories, with an average of 2.5 categories per image. Pascal-VOC is divided to a trainval set of 5,011 images and a test set of 4,952 images. Our training settings were identical to the ones used for MS-COCO. Notice that most previous works on Pascal-VOC used simple ImageNet pre-training, but some used additional data, like pre-training on MS-COCO or using NLP models like BERT. For a fair comparison, we present our results once with ImageNet pre-training, and once with additional pre-train data (MS-COCO) and compare them to the relevant works. Results appear in Table 6.

We can see from Table 6 that ASL achieves new state-of-the-art results on Pascal-VOC, with and without additional pre-training. In Table 8 in the appendix we compare different loss functions on Pascal-VOC, showing that ASL outperforms cross-entropy and focal loss.

4.1.3 NUS-WIDE

In appendix E we bring results on another common multi-label dataset, NUS-WIDE [7]. Table 9 shows that ASL again outperforms top previous approaches by a large margin, and reach new state-of-the-art result on NUS-WIDE.

Method	mAP (ImageNet Only Pretrain)	mAP (Extra Pretrain Data)
RNN [32]	91.9	-
FeV+LV [34]	92.0	-
SSGRL [4]	93.4	95.0
ML-GCN [6]	94.0	-
BMML [18]	-	95.0
ASL (ResNet101)	**94.4**	**95.3**
ASL (TResNet-L)	**94.6**	**95.8**

Table 6: **Comparison of ASL to known state-of-the-art models on Pascal-VOC dataset.** Metrics are in %.

4.1.4 Open Images

In appendix F we bring results on open images, a large scale dataset which consists of 9 million images. We can see from Table 11 that ASL significantly outperforms focal loss and cross-entropy, demonstrating that ASL is suitable for large datasets and extreme classification cases.

4.2. Additional Computer Vision Tasks

In addition to multi-label classification, we wanted to test ASL on other relevant computer vision tasks. Since fine-grain single-label classification and object detection tasks usually contain a large portion of background or long-tail cases [1, 16], and are known to benefit from using focal loss, we chose to test ASL on these tasks. In sections G and H in the appendix we show that ASL outperform focal loss on relevant datasets for these additional tasks, demonstrating that ASL is not limited to multi-label classification only.

5. Conclusion

In this paper, we present an asymmetric loss (ASL) for multi-label classification. ASL contains two complementary asymmetric mechanisms, which operate differently on positive and negative samples. By examining ASL derivatives, we gained a deeper understanding of the loss properties. Through network probability analysis, we demonstrate the effectiveness of ASL in balancing between negative and positive samples, and proposed an adaptive scheme that can dynamically adjusts the asymmetry levels throughout the training. Extensive experimental analysis shows that ASL outperforms common loss functions and previous state-of-the-art methods on popular multi-label classification benchmarks, including MS-COCO, Pascal-VOC, NUS-WIDE and Open Images.

References

[1] Peter L Bartlett and Marten H Wegkamp. Classification with a reject option using a hinge loss. *Journal of Machine Learning Research*, 9(Aug):1823–1840, 2008. 2, 3, 8

[2] Han Cai, Chuang Gan, Tianzhe Wang, Zhekai Zhang, and Song Han. Once for all: Train one network and specialize it for efficient deployment. In *International Conference on Learning Representations*, 2020. 7

[3] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 13

[4] Tianshui Chen, Muxin Xu, Xiaolu Hui, Hefeng Wu, and Liang Lin. Learning semantic-specific graph representation for multi-label image recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 522–531, 2019. 2, 8

[5] Zhao-Min Chen, Xiu-Shen Wei, Xin Jin, and Yanwen Guo. Multi-label image recognition with joint class-aware map disentangling and label correlation embedding. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 622–627. IEEE, 2019. 1, 7, 12

[6] Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo. Multi-label image recognition with graph convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5177–5186, 2019. 1, 2, 7, 8, 12

[7] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. Nus-wide: a real-world web image database from national university of singapore. In *Proceedings of the ACM international conference on image and video retrieval*, pages 1–9, 2009. 1, 8, 11

[8] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 113–123, 2019. 11

[9] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9268–9277, 2019. 3

[10] Thibaut Durand, Nazanin Mehrasa, and Greg Mori. Learning a deep convnet for multi-label classification with partial labels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 647–657, 2019. 1, 2, 4, 11

[11] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge 2007 (voc2007) results. 2007. 1, 8

[12] Bin-Bin Gao and Hong-Yu Zhou. Multi-label image recognition with multi-class attentional regions. *arXiv preprint arXiv:2007.01755*, 2020. 1, 2, 7, 12

[13] Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Time matters in regularizing deep networks: Weight decay and data augmentation affect early learning dynamics, matter little near convergence. *CoRR*, abs/1905.13277, 2019. 7

[14] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1314–1324, 2019. 11

[15] Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang. Learning deep representation for imbalanced classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5375–5384, 2016. 3

[16] kiat Chuan Tan. herbarium-2020-fgvc7, 2020. https://www.kaggle.com/c/herbarium-2020-fgvc7. 8, 12

[17] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Tom Duerig, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *arXiv preprint arXiv:1811.00982*, 2018. 1, 11

[18] Peng Li, Peng Chen, Yonghong Xie, and Dezheng Zhang. Bi-modal learning with channel-wise attention for multi-label image classification. *IEEE Access*, 8:9965–9977, 2020. 8

[19] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *CoRR*, abs/1708.02002, 2017. 2, 3, 13

[20] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2014. 1, 5, 7, 13

[21] Yongcheng Liu, Lu Sheng, Jing Shao, Junjie Yan, Shiming Xiang, and Chunhong Pan. Multi-label image classification via knowledge distillation from weakly-supervised detection. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 700–708, 2018. 2, 7, 11, 12

[22] Jinseok Nam, Eneldo Loza Mencía, Hyunwoo J Kim, and Johannes Fürnkranz. Maximizing subset accuracy with recurrent neural networks in multi-label classification. In *Advances in neural information processing systems*, pages 5413–5423, 2017. 1

[23] Kemal Oksuz, Baris Can Cam, Sinan Kalkan, and Emre Akbas. Imbalance problems in object detection: A review. *IEEE transactions on pattern analysis and machine intelligence*, 2020. 2, 6

[24] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses, 2021. 8

[25] Tal Ridnik, Hussam Lawen, Asaf Noy, Emanuel Ben Baruch, Gilad Sharir, and Itamar Friedman. Tresnet: High performance gpu-dedicated architecture. *arXiv preprint arXiv:2003.13630*, 2020. 7, 13

[26] Leslie N Smith. A disciplined approach to neural network hyper-parameters: Part 1–learning rate, batch size, momentum, and weight decay. *arXiv preprint arXiv:1803.09820*, 2018. 11

[27] Mingxing Tan and Quoc V Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *arXiv preprint arXiv:1905.11946*, 2019. 11

[28] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection, 2019. 13

[29] Jiang Wang, Yi Yang, Junhua Mao, Zhiheng Huang, Chang Huang, and Wei Xu. Cnn-rnn: A unified framework for multi-label image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2285–2294, 2016. 1

[30] Qian Wang, Ning Jia, and Toby P Breckon. A baseline for multi-label image classification using an ensemble of deep convolutional neural networks. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 644–648. IEEE, 2019. 11

[31] Ya Wang, Dongliang He, Fu Li, Xiang Long, Zhichao Zhou, Jinwen Ma, and Shilei Wen. Multi-label classification with label graph superimposing. *ArXiv*, abs/1911.09243, 2019. 1, 7

[32] Zhouxia Wang, Tianshui Chen, Guanbin Li, Ruijia Xu, and Liang Lin. Multi-label image recognition by recurrently discovering attentional regions. In *Proceedings of the IEEE international conference on computer vision*, pages 464–472, 2017. 1, 8

[33] Tong Wu, Qingqiu Huang, Ziwei Liu, Yu Wang, and Dahua Lin. Distribution-balanced loss for multi-label classification in long-tailed datasets. 2020. 1, 3

[34] Hao Yang, Joey Tianyi Zhou, Yu Zhang, Bin-Bin Gao, Jianxin Wu, and Jianfei Cai. Exploit bounding box annotations for multi-label object recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 280–288, 2016. 1, 8

[35] Jin Ye, Junjun He, Xiaojiang Peng, Wenhao Wu, and Yu Qiao. Attention-driven dynamic graph convolutional network for multi-label image recognition. 2019. 1

[36] Renchun You, Zhiyao Guo, Lei Cui, Xiang Long, Yingze Bao, and Shilei Wen. Cross-modality attention with semantic graph embedding for multi-label classification. In *AAAI*, pages 12709–12716, 2020. 1, 7, 12

[37] Sangdoo Yun, Seong Joon Oh, Byeongho Heo, Dongyoon Han, Junsuk Choe, and Sanghyuk Chun. Re-labeling imagenet: from single to multi-labels, from global to localized labels. *arXiv preprint arXiv:2101.05022*, 2021. 1

[38] Shifeng Zhang, Cheng Chi, Yongqiang Yao, Zhen Lei, and Stan Z. Li. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection, 2019. 7, 13

[39] Wenjie Zhang, Junchi Yan, Xiangfeng Wang, and Hongyuan Zha. Deep extreme multi-label learning. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, pages 100–107, 2018. 11

[40] Feng Zhu, Hongsheng Li, Wanli Ouyang, Nenghai Yu, and Xiaogang Wang. Learning spatial regularization with image-level supervisions for multi-label image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5513–5522, 2017. 12